254–256 Nucleic Acids Research, 1999, Vol. 27, No. 1 © 1999 Oxford University Press

BEST AVAILABLE COPY

SCOP: a Structural Classification of Proteins database

Tim J. P. Hubbard, Bart Ailey[1], Steven E. Brenner[3], Alexey G. Murzin[1] and Cyrus Chothia[2]

Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK, [1]Centre for Protein Engineering and [2]Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH, UK and [3]Department of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA

Received October 13, 1998; Accepted October 16, 1998

ABSTRACT

The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of all known proteins structures. The classification is on hierarchical levels: the first two levels, family and superfamily, describe near and far evolutionary relationships; the third, fold, describes geometrical relationships. The distinction between evolutionary relationships and those that aris from the physics and chemistry of proteins is a feature that is unique to this database, so far. The database can be used as a source of data to calibrate sequence search algorithms and for the generation of population statistics on protein structures. The database and its associated files are freely accessible from a number of WWW sites mirrored from URL http://scop.mrc-lmb.cam.ac.uk/scop/

INTRODUCTION

At present (October, 1998) the Brookhaven Protein Databank (PDB; 1) contains 7723 entries and the number is increasing by about 200 a month. These proteins have structural similarities with other proteins and, in many cases, share a common evolutionary origin. To facilitate access to this information, we have constructed the Structural Classification of Proteins (SCOP) database (2). It includes not only all proteins in the current version of the PDB, but many proteins for which there are published descriptions but whose co-ordinates are not yet available.

The classification of proteins in SCOP has been constructed by visual inspection and comparison of structures. Given the current limitations of purely automatic procedures, we believe this approach produces the most accurate and useful results. The unit of classification is usually the protein domain. Small proteins, and most of those of medium size, have a single domain and are, therefore, treated as a whole. The domains in large proteins are usually classified individually.

THE CLASSIFICATION

The classification of the proteins is on hierarchical levels.

Family

Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very similar; for example, globins with sequence identities of 15%.

Superfamily

Families, whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, the variable and constant domains of immunoglobulins.

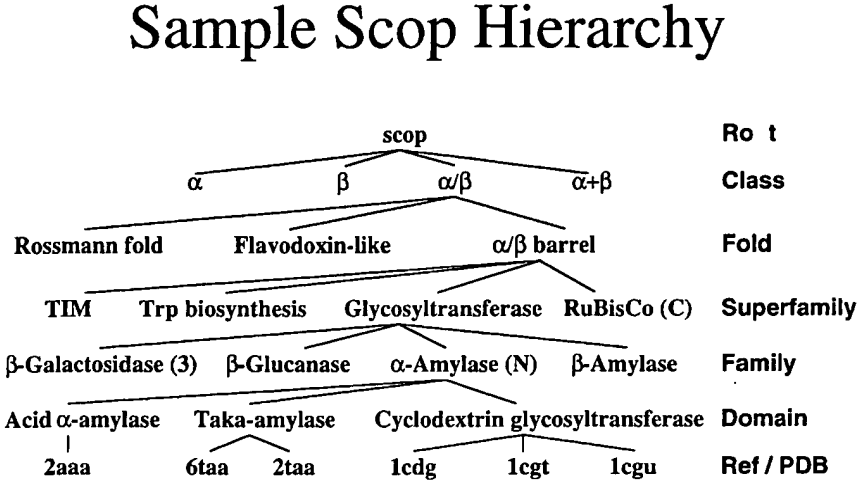
Common fold

Superfamilies and families are defined as having a common fold if their proteins have the same major secondary structures in the same arrangement and with the same topological connections (for recent reviews see refs 5 and 6). The structural similarities of proteins in the same fold category probably arise from the physics and chemistry of proteins favouring certain packing arrangements and chain topologies.

Class

The different folds have been grouped into classes. Most of the folds are assigned to one of the five structural classes: (i) all-α, those whose structure is essentially formed by helices; (ii) all-β, those whose structure is essentially formed by β-sheets; (iii) α/β, those with α-helices and β-strands; (iv) α+β, those in which α-helices and β-strands are largely segregated; and (v) multi-domain, those with domains of different fold and for which no homologues are known at present.

* To whom correspondence should be addressed. Tel: +44 1223 494983; Fax: +44 1223 494919; Email: th@sanger.ac.uk



BEST AVAILABLE COPY

Figure 1. Region of SCOP hierarchy. All the major levels, including class, fold, superfamily, and family are shown. Also shown are individual proteins and the lowest level, either the PDB coordinate identifier or a literature reference. Copyright © 1994 Steven E. Brenner; reproduced with permission.

Other classes have been assigned for peptides, small proteins, theoretical models, nucleic acids and carbohydrates. These hierarchical levels are illustrated in Figure 1.

There are now a number of other databases which classify protein structures, such as CATH (7,8), FSSP (9,10), Entrez (11) and DDBASE (12), however the distinction between evolutionary relationships and those that arise from the physics and chemistry of proteins is a feature that is, so far, unique to SCOP. Because functional similarity is implied by an evolutionary relationship but not necessarily by a physical relationship, we believe that this classification level is of considerable value, for example, as a way of reliably linking very distant sequence families.

ORGANISATION AND FACILITIES OF SCOP

The SCOP database is available as a set of tightly coupled hypertext pages on the world wide web (WWW) via URL: http://scop.mrc-lmb.cam.ac.uk/scop/

The interface to SCOP has been designed to facilitate both detailed searching of particular families and browsing of the whole database. To this end, there are a variety of different techniques for navigation as detailed below.

Browsing through the SCOP hierarchy. SCOP is organised as a tree structure. Entering at the top of the hierarchy, the user can navigate through the levels of Class, Fold, Superfamily, Family and Species to the leaves of the tree which are structural domains of individual PDB entries. An alternative hierarchy of Folds, Superfamilies and Families by the date of solution of the first representative structure is also provided.

From an amino acid sequence. The sequence similarity search facility allows any sequence of interest to be searched against databases of protein sequences classified in SCOP (see below) using the algorithms BLAST (13), FASTA or SSEARCH (14). SCOP can then be entered from the list of PDB chains found to be similar and the similarity can be displayed visually.

From a keyword. The keyword search facility returns a list of SCOP pages containing the word entered or combinations of words separated by a series of boolean operators.

From a PDB identifier. The PDB entry viewer links PDB entries to various graphical views, external databases and SCOP itself.

By history. Pages are provided that order folds, superfamilies and families by date of entry into PDB or publication. This is both for interest and to make it easier to keep up to date with the appearance of new folds or significant new members of existing folds.

In addition to the information on structural and evolutionary relationships contained within SCOP, each entry (for which co-ordinates are available) has links to images of the structure, interactive molecular viewers, the atomic co-ordinates, data on functional conformational changes, sequence data and homologues and MEDLINE abstracts.

To facilitate rapid and effective access to SCOP, a number of mirrors have been established, a full current list of which can be found via the above URL. The facilities provided by the various sites are always the same, so you will lose nothing by accessing your nearest mirror. The implementation does differ: for example, currently sequence similarity searching is always carried out at the main, scop.mrc-lmb.cam.ac.uk site, however, this is transparent to the user who will always be returned a search results page marked up with links to pages on the mirror that they started from.

OTHER USES OF SCOP

Evaluating the effectiveness of sequence alignment methods

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Despite this, the overall and relative capabilities of different search procedures have until recently been largely unknown. This is because it is difficult to verify algorithms on sample data as this requires large data sets of proteins whose evolutionary relationships are known

unambiguously and independently of the methods being evaluated (nearly all known homologs have been identified by sequence analysis, the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that, although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterised, or artificial test data (15).

As part of the maintenance of SCOP, new structures are automatically processed. One of the initial steps is to cluster the sequences of protein chains of known structures at different levels of sequence similarity. This has resulted in a series of non-redundant sequence databases, referred to as PDB40, PDB90, PDB95 (the number refers to percentage sequence identity as modified by the HSSP equation; 16). The chains chosen as representatives are those with the best structural 'quality' defined from an equation combining resolution, rfactor and procheck values (17). The final SCOP classification is used to annotate the headers of these fasta format files and to split them into domains. The result is a set of domain sequence databases, PDB40D, PDB90D, etc. where the full set of true and false pairwise relationships between the sequences can be inferred from the **scopcode** in the headers. These databases are used within SCOP for the sequence search facility (see above), however, they are also ideally suited as test data for the calibration of sequence searching algorithms. They have been used to calibrate the commonly used pairwise algorithms BLAST (13), WU-BLAST2 (18), FASTA and SSEARCH (14) (see ref. 15) as well as methods making use of multiple sequences such as Hidden Markov Models (19,20) and the recently developed iterative version of BLAST2 (21), referred to as psi-BLAST (22,23). The databases used for these studies are now freely available via the SCOP URL and can easily be filtered using the **scopcode** to extract subsets of sequences, e.g., to create a database with a single representative sequence for each fold, etc.

Statistics of protein structural data

With structural data conveniently organised into domains, it is straightforward to investigate the population statistics of the protein structures we currently know. A recent survey of the classification in SCOP (24) clearly shows that even after the high degree of redundancy in PDB has been taken into account, the frequency of occurrence of certain folds is much greater than would be expected by chance, as has been pointed out previously (25). The raw data needed to explore the classification in this way is provided in the form of the flat file from the SCOP URL.

CONCLUSIONS

We have found that the easy access to data and images provided by SCOP make it a powerful general-purpose interface to the PDB. The specific lower levels should be helpful for comparing individual structures with their evolutionary and structurally related counterparts. On a more general level, the highest levels of classification provide an excellent overview of the diversity of protein structures now known and would be appropriate both for researchers and students. Having created the classification we have found that it has many other uses, some of which have been listed here. We hope that other researchers will find yet more uses for the raw data files that are now provided with each release.

ACKNOWLEDGEMENTS

TJPH is grateful to the MRC/DTI/ZENECA LINK programme and AGM is grateful to the MRC for financial support.

REFERENCES

1 Abola,E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases: Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.
2 Murzin,A., Brenner,S.E., Hubbard,T.J.P. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
3 Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A. (1995) In Doolittle,R.F. (ed.), *Computer Methods for Macromolecular Sequence Analysis*. Academic Press, Orlando, FL.
4 Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A. and Chothia,C., *Acta Crystalog.*, in press.
5 Orengo,C. (1994) *Curr. Opin. Struct. Biol.*, **4**, 429–440.
6 Murzin,A.G. (1994) *Curr. Opin. Struc. Biol.*, **4**, 441–449.
7 Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) *Protein Engng*, **6**, 485–500.
8 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
9 Holm,L. and Sander,C. (1994) *Nucleic Acids Res.*, **22**, 3600–3609.
10 Holm,L. and Sander,C. (1996) *Science*, **273**, 595–602.
11 Hogue,C., Ohkawa,H. and Bryant,S.H. (1996) *Trends Biochem. Sci.*, **21**, 226–229.
12 Sowdhamini,R., Rufino,S.D. and Blundell,T.L. (1996) *Folding Design*, **1**, 209–220.
13 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
14 Pearson,W.R. (1996) *Methods Enzymol.*, **266**, 227–258.
15 Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
16 Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
17 Laskowski,R.A., Macarthur,M.W., Moss,D.S. and Thornton,J.M. (1993) *J. Appl. Crystallog.*, **26**, 283–291.
18 Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
19 Eddy,S.R. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.
20 Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D.J. (1994) *J. Mol. Biol.*, **235**, 1501–1531.
21 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
22 Park,J.H., Teichmann,S.A., Hubbard,T.J. and Chothia,C. (1997) *J. Mol. Biol.*, **273**, 349–354,
23 Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C., *J. Mol. Biol.*, in press.
24 Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1997) *Curr. Opin. Struct. Biol.*, **7**, 369–376.
25 Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) *Nature*, **372**, 631–634.

BEST AVAILABLE COPY